# Gender Classification by Deep Learning on Millions of Weakly Labelled Images

Sen Jia, Thomas Lansdall-Welfare and Nello Cristianini
Intelligent Systems Laboratory
University of Bristol
Bristol
United Kingdom
Email: sen.jia@bris.ac.uk,thomas.lansdall-welfare@bris.ac.uk,nello.cristianini@bris.ac.uk

***Abstract*—When analysing human activities using data mining or machine learning techniques, it can be useful to infer properties such as the gender or age of the people involved. This paper focuses on the sub-problem of gender recognition, which has been studied extensively in the literature, with two main problems remaining unsolved: how to improve the accuracy on real-world face images, and how to generalise the models to perform well on new datasets. We address these problems by collecting five million weakly labelled face images, and performing three different experiments, investigating: the performance difference between convolutional neural networks (CNNs) of differing depths and a support vector machine approach using local binary pattern features on the same training data; the effect of contextual information on classification accuracy; and the ability of convolutional neural networks and large amounts of training data to generalise to cross-database classification. We report record-breaking results on both the Labeled Faces in the Wild (LFW) dataset, achieving an accuracy of 98.90%, and the Images of Groups (GROUPS) dataset, achieving an accuracy of 91.34% for cross-database gender classification.**

***Index Terms*—Gender Classification; Deep Learning; Big Data;**

## I. Introduction

The classic task of face gender recognition has recently attracted new attention, mostly due to the availability of large sets of images collected "in the wild". Applications are readily found in many areas, for example in the analysis of gender bias in news media content [1], [2]. The emphasis of this new phase of research is on avoiding images collected under controlled conditions (e.g. in background, pose or illumination), and focusing efforts on the more challenging case of natural images.

While significant results have been obtained with the traditional two-step procedure (feature extraction followed by statistical classifiers) [3], the introduction of CNNs has recently led to further improvements in performance [4], [5]. CNNs are capable of learning their own features, at the same time as learning the classifier, but to do so they require the tuning of an enormous number of free parameters, and hence the availability of very large training sets to avoid over-fitting.

This leads to two separate challenges: the use of efficient hardware, such as Graphics Processing Units (GPUs) to train the networks, and the creation of very large sets of images, obtained in uncontrolled conditions, that are labelled by gender. As the required number of images can range into the millions, the use of hand-applied labels is not an option. If these problems can be solved, however, CNNs hold the promise of leading to significant performance improvements [6], due mostly to their capability to essentially design their own features and their ability to generalise well to new scenarios.

The same situation can be seen in different image classification tasks, for example, in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [7], where millions of images were used to train a deep classifier to solve a real-world vision task. For face recognition, big data is also a key factor in obtaining human-level performance [8].

In this paper, we investigate the effects of combining the power of CNNs with the information contained in massive, weakly labelled datasets, collected from the web. We collect five million publicly available face images that are weakly labelled using gender-specific queries derived from the Internet Movie Database (IMDB), following the procedure in [3]. These images are used as our training data for three different experiments, reporting performance on the LFW dataset [9] and the GROUPS dataset [10] for comparability.

The first experiment is aimed at investigating the difference in performance for varying depths of CNN, and comparing directly with a Support Vector Machine (SVM) using Local Binary Pattern (LBP) features approach using the same training and test sets. We show that convolutional neural networks can outperform traditional approaches (e.g. [3]), while the use of deeper networks allows for large sets of noiser training data to be used to increase performance (e.g. deeper networks plus weakly labelled data can improve on previously reported CNN approaches [4]).

In the second experiment we investigate using a larger bounding box for the face region, on exactly the same data as the first experiment, and show that we can further increase performance using the additional contextual information that the larger bounding box contains. We show that this change leads to an improvement of $0.65\%$ over the best result in the first experiment, with a record-breaking accuracy of $98.90\%$ on LFW.

Finally, in the third experiment, we use a recently proposed face detector [11] to replace the Voila Jones (VJ) algorithm. We apply this to the weakly labelled images collected from

the web, generating five million training images, and test our performance on the unaligned version of the LFW dataset, along with the GROUPS dataset. This introduces a more difficult scenario, where the images are unaligned and in a cross-database setting. We show that even in this more difficult setting, we can achieve an accuracy of $98.69\%$ on the unaligned LFW and $91.34\%$ on the GROUPS dataset.

To summarize our work, the contributions include:

- We directly compare CNNs of varying depth with previous work, showing that on the same training data CNNs out-perform SVMs using LBP features, and that increasing the depth of the network allows for large sets of noiser training data to improve performance over shallower networks.
- We show that including more contextual information about the face region, by using a larger bounding box for the face region will improve performance, leading to a record-breaking accuracy of $98.90\%$ on the LFW dataset.
- We introduce a more difficult setting, where the test data is unaligned and in a cross-database setting, making it more comparable with real-world images. We report the best cross-database accuracy of $98.69\%$ on the unaligned LFW and $91.34\%$ on the GROUPS dataset, showing that our trained network can generalise to new settings.

The rest of the paper is organized as follows: Section II covers a literature review of face gender recognition; Section III details the three experiments, their settings and our results; while Section IV concludes the paper.

## II. Related Work

Face analysis under controlled conditions has been well studied, but performing gender recognition in unconstrained environments is still a difficult task in computer vision. Differences in reported performances between testing in cross-database (different sources for training and test data) and within-database (same source) settings further highlights the need for methods that generalise well across different datasets.

Shan [12] manually chose $7{,}443$ face images from LFW where difficult images, such as profile, rotated and baby faces, were not considered. Multi-scale LBP features were extracted before the $500$ most discriminative bins were selected using Adaboost. He achieved $94.81\%$ on a subset of LFW using SVM with 5-fold cross validation in a within-database setting. Similarly, Ren [13] achieved $98\%$ on a smaller subset of $6{,}840$ faces from LFW, using three different types of features. Scale-Invariant Feature Transform (SIFT), Histograms of Oriented Gradients (HOG) and Gabor wavelet features were combined, before applying RealAdaboost with a complexity penalty term to choose useful features whilst easing the computational cost. However, the results of Shan [12] and Ren [13] were both reported on easy, subsamples of LFW.

For the GROUPS dataset [10] in a within-database setting, Fazl-Ersi [14] combined SIFT, LBP and Colour Histogram (CH) and achieved an accuracy of $91.59\%$. Bekios-Calfa [15] proposed to recognize face gender using facial attribute dependencies, such as age and pose, achieving $80.5\%$ on the subset of GROUPS without children. Han [16] used biologically inspired features (BIF) and SVM on all the faces from GROUPS and achieved an accuracy of $87.1\%$.

It is worth noting that there is a performance drop when training and test on different datasets. In a cross-database setting, where methods are trained on data from a different source to the test set, Pablo Dago-Casas [17] reported an accuracy of $81.02\%$ on a subset of $14{,}760$ face images from the GROUPS dataset where low resolution faces had been removed, along with randomly removing a number of male faces to achieve an equal gender distribution. This was achieved using raw pixel values, LBP and Gabor jets features and then applying Principal Component Analysis (PCA) for dimension reduction. They also reported $89.77\%$ on the LFW database in a cross-database test.

Mansanet [18] applied the Sobel filter and low-pass filter to detect important patches from face images which were then used to train a deep neural network. The best results they achieved on GROUPS and LFW were $83.03\%$ and $94.48\%$ respectively.

Jia [3] created an automatic gender classification system using four million weakly labelled face images collected from the web. They report an accuracy of $96.86\%$ on $10{,}147$ face images detected using the VJ algorithm from LFW, obtained with a large margin linear classifier using approximately 60,000 multi-scale LBP features.

Recently, Antipov [4] simplified the CNN architecture used in object classification [19] for gender recognition. In their work, they varied the depth of CNN and found that shallower CNNs can be trained efficiently without much performance loss. Their ensemble of CNNs gave an accuracy of $97.31\%$ on the same LFW test set used in [3], training on $494{,}414$ faces images from the CASIA database [20].

Castrillon [5] combined CNNs with more types of vision features from different facial regions. They achieved the highest accuracy on the LFW and GROUPS in a cross-database setting, $98\%$ and $90.14\%$ respectively.

## III. Experiments

In this paper, we perform three different experiments to investigate the difference in performance between CNNs of varying depths and a SVM approach using LBP features [3] on the same training and test datasets; the effect of contextual information on classification accuracy by using a larger bounding box; and the ability of CNNs trained with large amounts of noisy data to generalise well in cross-database classification.

### A. Weakly labelled data

For the training data in every experiment, we trained on millions of publicly available images retrieved from the web. To retrieve these images, we first generated a list of gender-specific queries from IMDB. The list contained equal numbers of male and female names, such as John (male) and Mary (female). Using this list, we then queried search engines to retrieve sets of images, following the protocol in [3]. Using this approach, we collected a total of $6{,}727{,}509$ images, using

TABLE I
RELATED WORK AND THEIR REPORTED ACCURACIES.

| Authors | Test set | Features | Classifiers | Result | Cross |
|---|---|---|---|---|---|
| Shan [12] | LFW(7,443) | Boosted LBP | SVM | 94.81% | N |
| Ren [13] | LFW(6,840) | HOG<br>SIFT<br>Gabor | SVM | 98% | N |
| Fazl-Ersi [14] | GROUPS(14,760) | LBP<br>SIFT<br>CH | SVM | 91.59% | N |
| Bekios-Calfa [15] | GROUPS(22,948) | Pixel | LDA | 80.5% | N |
| Han [16] | GROUPS(28,231) | BIF | SVM | 87.1% | N |
| Dago-Casas [17] | GROUPS(14,760)<br>LFW(13,088) | LBP<br>PCA | LDA | 81.02%<br>89.77% | Y |
| Mansanet [18] | GROUPS(14,760)<br>LFW(13,233) | DNN | Class-posterior | 83.03%<br>94.48% | Y |
| Jia [3] | LFW(10,147) | Multi-scale<br>LBP | C-Pegasos | 96.86% | Y |
| Antipov [4] | LFW(10,147) | CNN | Softmax<br>(Ensemble) | 97.31% | Y |
| Castrillon [5] | GROUPS(28,231)<br>LFW(13,233) | LBP<br>HOG<br>LOSIB<br>CNN | SVM | 90.14%<br>98.00% | Y |
| **Ours** | LFW(10,147) | CNN | Softmax | 98.90% | Y |
| **Ours** | LFW(13,061) | CNN | Softmax | 98.69% | Y |
| **Ours** | GROUPS(24,743) | CNN | Softmax | 96.10% | Y |
| **Ours** | GROUPS(28,163) | CNN | Softmax | 91.34% | Y |

150,000 names from each gender class. Each of these images is further processed to detect if it is a face image using different approaches in each experiment.

### *B. Experiment 1: Effect of Network Depth*

In Experiment 1, we investigated what effect the depth of the convolutional neural network had on gender classification accuracy using weakly labelled data, keeping a fixed training and test set.

*1) Training set:* We used the VJ algorithm on the weakly labelled data to first extract a facial bounding box, before determining if a face was present by extracting facial landmarks, resulting in 4,227,792 face images, following the procedure in [3].

*2) Network architectures:* We used the deep learning library Caffe [21] to create three different networks of varying depth by tuning the VGG-16 model suggested by Parkhi [22]. This network was first proposed by Simonyan [19] for object classification, achieving state-of-the-art results in the ILSVRC. Previously, in [4], they used a simplified version of the VGG model with 13 layers as their starting CNN, before simplifying it further to a 6-layer architecture.

We started from a similar 6-layer architecture to [4], which has the same number of layers but with more filters at each convolutional layer, which we name G6. This network was trained from scratch using a batch size of 512. We trained for 300,000 iterations, such that the model was trained for $512 \times 300{,}000 \div 4{,}227{,}792 \approx 36.33$ epochs. The initial learning rate was set to 0.01 and reduced by 10 times after every 100,000 iterations. The momentum coefficient was 0.9 and the weight decay was 0.0005. The dropout rate was set to 0.5 and we saved a snapshot of the model after every 10,000 iterations. The model was trained on small face images of $40 \times 40$ pixels before taking crops from the four corners and centre, for inputs of size $32 \times 32$.

Next, we add four and ten more convolutional layers to the G6 architecture, creating the G10 and G16 networks respectively, where the G16 has the same archteciture as VGG-16 [19]. These two architectures were fine-tuned from the pre-trained Parkhi model[1] [22]. These were fine-tuned using larger faces than the G6, with images of $256 \times 256$ pixels cropped from the corners and centre to a size of $224 \times 224$ for the input layer.

We set a mini-batch size of 128 for G10 and the training iterations to 200,000, giving us 6.05 epochs. The learning rate was 0.001 at the beginning, decreasing by 10 times after every 80,000 iterations. For the G16, the mini-batch size was further reduced to 64 because of our GPU memory. We increased the training iterations to 350,000, giving us around 5.3 epochs. The learning rate reduced by 10 times after every 150,000 iterations.

The G16 is the deepest model used in our work due to the memory size of available GPUs (NVIDIA Titan X, 12 gigabytes). Each of the network architectures used are detailed in Table II where $64@3 \times 3$ denotes 64 convolutional filters with a size of 3 by 3. Every convolutional and fully connected layer is followed by a Rectified Linear Unit (ReLU). Dropout layers are only used after each fully connected layer. Finally, a fully connected layer with two neurons is added as the output layer for binary classification.

[1]Available from: http://www.robots.ox.ac.uk/~vgg/software/vgg_face/

TABLE II
CNN ARCHITECTURES OF DEPTH 6, 10 AND 16.

| Shallow (G6) | Medium (G10) | Deep (G16) |
|---|---|---|
| Input: $32 \times 32$ | Input: $224 \times 224$ | Input: $224 \times 224$ |
| Conv: $64@3 \times 3$ | Conv: $64@3 \times 3$ | Conv: $64@3 \times 3$ |
| Conv: $64@3 \times 3$ | Maxpool: $2 \times 2$ | Conv: $64@3 \times 3$ |
| Maxpool: $2 \times 2$ | Conv: $128@3 \times 3$ | Maxpool: $2 \times 2$ |
| Conv: $128@3 \times 3$ | Maxpool: $2 \times 2$ | Conv: $128@3 \times 3$ |
| Conv: $128@3 \times 3$ | Conv: $256@3 \times 3$ | Conv: $128@3 \times 3$ |
| Maxpool: $2 \times 2$ | Maxpool: $2 \times 2$ | Maxpool: $2 \times 2$ |
| | Conv: $512@3 \times 3$ | Conv: $256@3 \times 3$ |
| | Conv: $512@3 \times 3$ | Conv: $256@3 \times 3$ |
| | Maxpool: $2 \times 2$ | Conv: $256@3 \times 3$ |
| | Conv: $512@3 \times 3$ | Maxpool: $2 \times 2$ |
| | Conv: $512@3 \times 3$ | Conv: $512@3 \times 3$ |
| | Maxpool: $2 \times 2$ | Conv: $512@3 \times 3$ |
| | | Conv: $512@3 \times 3$ |
| | | Maxpool: $2 \times 2$ |
| | | Conv: $512@3 \times 3$ |
| | | Conv: $512@3 \times 3$ |
| | | Conv: $512@3 \times 3$ |
| | | Maxpool: $2 \times 2$ |
| Fully connected:512 | Fully connected:4096 | Fully connected:4096 |
| | Fully connected:4096 | Fully connected:4096 |

TABLE III
RESULTS FROM EXPERIMENT 1 AND 2, SHOWING THAT OUR G16 NETWORK SETS A NEW RECORD FOR ACCURACY ON THE SAME LFW SUBSET USED IN [3], [4], AND THAT INCREASING THE BOUNDING BOX SIZE IN EXPERIMENT 2 FURTHER IMPROVES PERFORMANCE.

| | G6 | G10 | G16 | [3] | [4] |
|---|---|---|---|---|---|
| Experiment 1 | 95.80% | 96.94% | **98.25%** | 96.86% | 97.31% |
| Experiment 2 | 96.67% | 97.79% | **98.90%** | 96.86% | 97.31% |

*3) Test set:* Each of the networks in this experiment was tested on the subset of 10,147 face images from LFW reported in [3], [4], to allow for comparability. The subset of face images is obtained by running the VJ algorithm on all images in LFW and removing those that do not contain a face, or any facial landmarks. The test set is further split into two subsets, used as a validation and test set, which are then reversed with the experiment conducted again, using the same testing protocol as [3]. Accuracies are reported as the average over the two-fold cross-validation. We ensure that there is no subject intersection between the training and test set, or the validation and test sets by removing near duplicates based upon pixel-wise comparisons and their query names, following the procedure in [8].

*4) Results:* In Table III, we compare our results from the three varying depths of network with the results reported for the same LFW test set in [3] and [4]. We can see that in Experiment 1, as more layers are added, our accuracy on the test set increases, achieving an accuracy of 98.25% using the G16 network.

### C. Experiment 2: Effect of Bounding Box Size

In Experiment 2, we investigated what effect the size of the facial bounding box had on gender classification accuracy, comparing all three architectures from Experiment 1 on the same fixed training and test set. This was motivated by research by Kumar [23] showing that humans can correctly verify with an accuracy of 94.27% if two images contain the same person even when a tightly cropped bounding box of the face has been removed, suggesting that there may be additional information we can use for gender classification by using a larger bounding box (see Fig. 1).

*1) Training set:* We used the same VJ algorithm on the weakly labelled data to first extract a facial bounding box, before determining if a face was present by extracting facial landmarks, resulting in 4,227,792 face images, following the procedure in [3]. Further to this, we then increased the area within the facial bounding box of each image by 1.5 times, allowing for more of the contextual information around the face to be included within our training images.

*2) Network architectures:* For this experiment, we used exactly the same network architectures as detailed in Experiment 1, changing only the size of the bounding boxes for the training set. The size of the images to the input layers for each network did not change, with each image being scaled down and cropped to $32 \times 32$ for the G6 network, and $224 \times 224$ for the G10 and G16 networks as before. All network parameters were kept the same as Experiment 1.

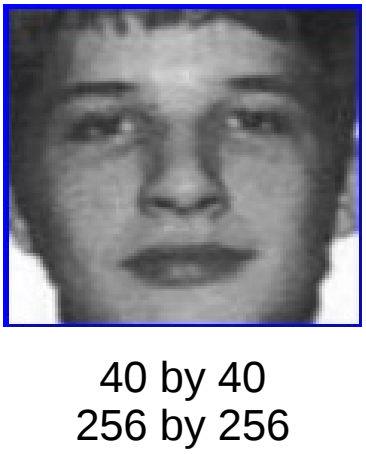
*3) Test set:* We tested each of the three networks trained on images with larger bounding box areas on the same set of images as Experiment 1, with the larger bounding box also used when extracting facial bounding boxes from the test images. This setting remains directly comparable with previous works [3], [4] on the same subset of LFW.

*4) Results:* In Table III, we also compare our results from Experiment 2 for the three varying depths of network. The accuracies reported are tested again on the same LFW test set in [3] and [4]. We can see that similarly to Experiment 1, we find that as more layers are added, our accuracy on the test set increases, achieving a record-breaking accuracy of 98.90% using the G16 network, an improvement over the same network with a smaller bounding box in Experiment 1. This suggests that using a less tightly cropped face image for training in gender classification can lead to small performance improvements by incorporating more contextual information.

### D. Experiment 3: A More Challenging Setting

In Experiment 3, we wished to test how well our networks perform in a more challenging setting, where the face images in the test set are not aligned, and with more challenging faces being kept (e.g. we do not require facial landmarks to be present), resulting in a larger, more difficult test set. Additionally, we wished to test our performance on the GROUPS dataset [10], a set of 28,231 faces annotated with age, gender and location information from a collection of 5,080 group photos. For this experiment, we used the best settings we had found from the previous two experiments, namely we used the G16 network with larger bounding box areas.

*1) Training set:* For this experiment, we used the recently proposed face detector [11] on the weakly labelled data, replacing the VJ algorithm used in the first two experiments.



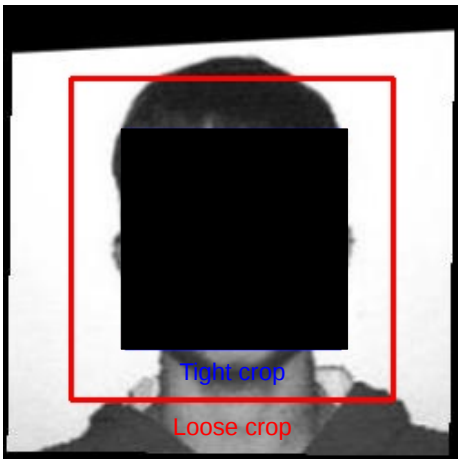




Fig. 1. An example of a tightly cropped face from the LFW dataset, along with the additional information we can capture by using a larger bounding box. (left, each face image is resized to 40 by 40 and 256 by 256 pixels for different experiment settings)

This was performed since the new face detector has been shown to outperform the VJ algorithm, increasing the number of face detections on the weakly labelled data, therefore allowing for more images to be kept in the training set. In total, we used 5,286,130 face images from the original weakly labelled data in Section III-A for training our networks in this experiment.

*2) Network architecture:* We wished to only apply our best architecture found in the previous experiments to the new, more challenging problem setting. As such, we used the G16 architecture detailed in Experiment 1 trained using the weakly labelled data with a different face detector applied, and the larger bounding boxes from Experiment 2. All other configurations for the network were kept fixed.

*3) Test set:* In this experiment, we tested our trained network on two different datasets, the unaligned version of the LFW dataset, and the GROUPS dataset. For the unaligned version of LFW, we applied the new face detector [11] to the test images, obtaining a test set of 13,061 loosely cropped face images, containing more difficult samples than usually reported. For the publicly available GROUPS dataset, we first identified all the faces from the images using the given eye coordinates, resulting in 28,231 face images, the full dataset as reported in the literature (GROUPS-all). When testing on each of these test sets, we used the other test set as a validation set (i.e. LFW was used as a validation set when testing on GROUPS-all, and vice versa). We further considered two different subsets of the GROUPS data reported in the literature to allow us to compare with previously published results, namely all face images with an interocular distance greater than or equal to 20 pixels (GROUPS-A, 16,368 images), and all face images where the person was annotated as being older than 12 (GROUPS-B, 24,743 images).

TABLE IV
RESULTS FOR EXPERIMENT 3 IN THE MORE CHALLENGING SETTING, USING THE UNALIGNED LFW AND GROUPS TEST SETS IN A CROSS-DATABASE SETTING.

| Test set | G16 | [5] | [18] | [15] |
|---|---|---|---|---|
| Unaligned LFW | **98.69%** | 98.00% | - | |
| GROUPS-all | **91.34%** | 90.14% | - | - |
| GROUPS-A | **91.80%** | - | 83.03% | - |
| GROUPS-B | **96.10%** | - | - | 80.53% |

*4) Results:* Table IV shows the performance of our G16 network in the new, more challenging setting we propose, alongside a comparison with the previous best results in the literature, to the best of our knowledge. We can see that in this cross-database setting, we significantly out-perform previous methods on the GROUPS-A and GROUPS-B test sets, with slightly increased performance over the current best result for the Unaligned LFW and GROUPS-all test sets. It should be noted that the result from [5] is reported on the entire LFW dataset, so is not directly comparable, although it is the closest result we have found to the new, more challenging setting.

## IV. CONCLUSIONS

The task of automatically classifying the gender of a person, based on an image of their face, has been investigated under different conditions. In this study, we conducted several experiments to determine the effects that different design choices

can have on the performance of a gender classification system working on images taken under uncontrolled conditions.

Among our contributions has been to show the effect of training a deep convolutional network on a vast, weakly labelled training set, formed of millions of face images, collected from the web. Our results are compared with previous studies on the same test sets, which were either performed with simpler algorithms on similar training sets, or with similar algorithms on smaller training sets. We achieved the best performance ever reported on the LFW test set used in [3], [4], moving the best attainable accuracy from $97.31\%$ to $98.90\%$.

More importantly, we have investigated the effect of network depth on the classification performance, finding that for our experiments using weakly labelled data, a deeper CNN out-performed a shallower one with the same training data; that using a larger bounding box for the face region can improve performance; and that our proposed CNN generalises well to other test sets, achieving a state-of-the-art performance of $91.34\%$ on the GROUPS-all test set. In this setting, we demonstrate that while we achieve further improvements over those previously reported, that there is still space for increasing the accuracy in face gender classification tasks.

## ACKNOWLEDGMENT

The Titan X graphic card used for this research was donated by the NVIDIA Corporation.

## REFERENCES

[1] S. Jia, T. Lansdall-Welfare, and N. Cristianini, "Measuring gender bias in news images," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15 Companion. New York, NY, USA: ACM, 2015, pp. 893–898. [Online]. Available: http://doi.acm.org/10.1145/2740908.2742007

[2] S. Jia, T. Lansdall-Welfare, S. Sudhahar, C. Carter, and N. Cristianini, "Women are seen more than heard in online newspapers," *PLoS ONE*, vol. 11, no. 2, p. e0148434, 02 2016. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0148434

[3] S. Jia and N. Cristianini, "Learning to classify gender from four million images," *Pattern Recogn. Lett.*, vol. 58, no. C, pp. 35–41, Jun. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2015.02.006

[4] G. Antipov, S. A. Berrani, and J. L. Dugelay, "Minimalistic cnn-based ensemble model for gender prediction from face images," *Pattern Recognition Letters*, 2015.

[5] M. C. Santana, J. Lorenzo-Navarro, and E. Ramón-Balmaseda, "Descripors and regions of interest fusion for gender classification in the wild," *CoRR*, vol. abs/1507.06838, 2016. [Online]. Available: http://arxiv.org/abs/1507.06838

[6] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *Intelligent Systems, IEEE*, vol. 24, no. 2, pp. 8–12, 2009.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[8] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[10] A. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. CVPR*, 2009.

[11] M. Uřičář, V. Franc, D. Thomas, S. Akihiro, and V. Hlaváč, "Real-time Multi-view Facial Landmark Detector Learned by the Structured Output SVM," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015*, vol. 02, May 2015, pp. 1–8.

[12] C. Shan, "Learning local binary patterns for gender classification on real-world face images," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 431–437, 2012.

[13] H. Ren and Z.-N. Li, "Gender recognition using complexity-aware local features," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, Aug 2014, pp. 2389–2394.

[14] E. Fazl-Ersi, M. E. Mousa-Pasandi, R. Laganiere, and M. Awad, "Age and gender recognition using informative features of various types," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5891–5895.

[15] J. Bekios-Calfa, J. M. Buenaposada, and L. Baumela, "Robust gender recognition by exploiting facial attributes dependencies," *Pattern Recognition Letters*, vol. 36, pp. 228 – 234, 2014.

[16] H. Han and A. K. Jain, "Age, gender and race estimation from unconstrained face images," *MSU Technical Report, MSU-CSE-14-5*, pp. 1–9, 2014.

[17] P. Dago-Casas, D. Gonzalez-Jimenez, L. L. Yu, and J. Alba-Castro, "Single- and cross- database benchmarks for gender classification under unconstrained settings," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, Nov 2011, pp. 2152–2159.

[18] J. Mansanet, A. Albiol, and R. Paredes, "Local deep neural networks for gender recognition," *Pattern Recognition Letters*, vol. 70, pp. 80 – 86, 2016.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Eprint Arxiv*, 2014.

[20] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *Eprint Arxiv*, 2014.

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[22] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *Proceedings of the British Machine Vision*, vol. 1, no. 3, p. 6, 2015.

[23] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Describable visual attributes for face verification and image search," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 10, pp. 1962–1977, Oct 2011.